Lecture 8: Learning Theory and Concentration

Scribe: Yatin Dandi Lecture: Rajat Mittal

IIT Kanpur

These notes are based on Chapter 3 of O'Donnell [2014].

1 Complexity of Boolean Functions in terms of Fourier Spectrum

In the past lectures, we have studied various measures of complexity of boolean functions such as degree and total influence. Our first objective in the present lecture is to characterize the complexity of boolean functions in terms of their Fourier spectrum. We recall that for boolean functions f having range $\{-1, 1\}$, Parseval's identity implies that the sum over squared fourier coefficients is 1 i.e $\sum\limits_{S\subseteq[n]} \widehat{f}^2(S) = 1$. Thus $\widehat{f}^2(S)$ can be viewed as defining a discrete probability distribution over the characters or equivalently over the sets $S \subseteq [n]$. Using the above distribution, the complexity of a function f can be characterized in terms of the probability mass situated on high degree characters [O'Donnell, 2014]:

Definition 1. *We say that the Fourier spectrum of $f\colon \{-1,1\}^n \to \mathbb{R}$ is ϵ-concentrated on degree up to k if*

$$\sum_{S\in[n],|S|>k} \hat{f}(S)^2 \leq \epsilon$$

The definition indicates that the name ϵ-concentrated is slightly unnatural, since the weight on terms upto degree k is actually $1-\epsilon$. For instance a function is 0-concentrated on degree up to k if and only if $\deg f \leq k$. Using a generalization of the earlier proof of the inequality $\deg f > \frac{\log n}{\log\log n}$, we can show that a function f satisfying $\leq d$ depends on atmost $d2^{d-1}$ variables.

The weight $1-\epsilon$ also corresponds to the probability mass over the characters upto degree k using the probability distribution defined above.

Recall that the degree of a boolean function only depends on the largest degree term having a non-zero fourier coefficient and disregards the size of the fourier coefficients. Thus the degree of a boolean function is discontinous in the sense that it be large even for functions having arbitrarily small coefficients for the highest degree terms. The above definition of ϵ-concentration yields a notion of the approximate degree of a function that is more robust and continous, by defining $\deg'_\epsilon f$ to be the minimum degree d such that f is ϵ-concentrated on degree up to d. In the last lecture, we studied another notion of approximate degree, defined as:

$$\widetilde{\deg}_\epsilon(f) = \min_{\substack{\text{polynomial } p:\\ |p(x)-f(x)|\leqslant\epsilon \ \forall x}} \deg(p)$$

The above definition is equivalent to saying that $\widetilde{\deg}_\epsilon(f)$ is the minimum degree d such that a d degree polynomial $p(x)$ is close to $f(x)$ in the ∞ norm i.e $\|p(x)-f(x)\|_\infty \leq \epsilon$ where $\|p(x)-f(x)\|_\infty = \max_x |p(x)-f(x)|$ However, using Parseval's identity, we know that $\|p(x)-f(x)\|_2 = \left\|\hat{p}(x)-\hat{f}(x)\right\|_2$ where the norm corresponds to the 2-norm. Thus the above two notions of approximate degree differ only in the chosen norm.

1.1 Relation with Influence

Recall that the total influence of any function can by bounded by the degree i.e. $Inf(f) \leq \deg(f)$. Thus a function with a small degree must have small total influence. However, as can be seen from the example of the OR function, the converse is false. The OR function has degree n whereas the influence of each variable is $\frac{1}{2^{n-1}}$. Thus the total influence $\frac{n}{2^{n-1}}$ for the OR function is small, even though it has the maximum possible degree.

However, an application of Markov's inequality reveals that the total influence bounds the degree in an ϵ-concentrated sense. Concretely, observe that total influence $\mathit{Inf}(f)$ equals the expected degree of the function f under the probability distribution defined using the fourier coefficients i.e $\mathit{Inf}(f) = \sum_{S \in [n]} |S| \hat{f}^2(S)$. Thus applying Markov's inequality to the non-negative random variable $|S|$ yields:

$$\Pr_{S \sim S_f}[|S| \geq k] \leq \frac{\mathit{Inf}(f)}{k}$$

, where S_f is a random subset sampled with probabilities $\Pr[S_f = S] = \hat{f}^2(S)$. Choosing $k = \frac{\mathit{Inf}(f)}{\epsilon}$ gives:

$$\Pr_{S \sim S_f}[|S| \geq \frac{\mathit{Inf}(f)}{\epsilon}] \leq \epsilon.$$

Thus a function f with influence $\mathit{Inf}(f)$, is ϵ-concentrated upto degree $\frac{\mathit{Inf}(f)}{\epsilon}$.

1.2 Relaxation of ϵ-concentration

We can further generalize the notion of ϵ-concentration to arbitrary collections of subsets instead of upto k degree terms. We obtain the following notion:

Definition 2. *Let $\mathcal{F}$ be a collection of subsets $S \subseteq [n]$. We say that the Fourier spectrum of $f : \{-1,1\}^n \to \mathbb{R}$ is ϵ-concentrated on $\mathcal{F}$ if*

$$\sum_{\substack{S \subseteq [n] \\ S \notin \mathcal{F}}} \widehat{f}(S)^2 \leq \epsilon$$

Equivalently, we can say that the Fourier spectrum of $f : \{-1,1\}^n \to \mathbb{R}$ is ϵ-concentrated on $\mathcal{F}$ if $\Pr_{S \sim S_f}[S \notin \mathcal{F}] \leq \epsilon$

In the following section, we will relate the above notion to the learnability of different classes of functions.

2 Learning Theory

Machine learning is an umbrella-term encapsulating algorithms that "learn" functions, probability distributions or tasks defined on an input space, using the values of the function on only a finite number of training points. The training points can be viewed as arising from an oracle. Thus the goal of machine learning is to learn the underlying functions using a minimum number of queries to the oracle.

Learning theory provides a framework for understanding which classes of functions, or hypotheses can be learned using a finite or a small number of training points.

For functions defined on the boolean hypercube, any function f can be uniquely identified through its values at the 2^n input points. However, as mentioned above, it is desirable to instead learn such functions "efficiently", i.e. using a minimum number of queries. For arbitrary functions, an adversary argument shows that learning without quering the values at all the points is infeasible. However, we will show that when functions are known to belong to a given a class of functions, denoted by **concept class**, learning can be made more efficient.

The function's value at the training points can be obtained in one of the following two ways:

1. Query model: The algorithm can choose the points where the function's values are obtained.
2. Random model: The points where the function's value is available are sampled randomly.

In this lecture, we will analyze the random model. We'll further assume that the points are sampled uniformly on $\{-1,1\}$. Concretely, we adopt the following notion of learning, known as Probably Approximately Correct (PAC) learning [Valiant, 1984, O'Donnell, 2014]:

Definition 3. *In the model of PAC ("Probably Approximately Correct") learning under the uniform distribution on $\{-1,1\}^n$, a learning problem is identified with a concept class $\mathcal{C}$, which is just a collection of functions $f : \{-1,1\}^n \to \{-1,1\}$. A learning algorithm A for $\mathcal{C}$ is a randomized algorithm which has limited access to an unknown target function $f \in \mathcal{C}$.*

In addition, A is given as input an accuracy parameter $\epsilon \in [0,1/2]$. The output of A is required to be (the circuit representation of) a hypothesis function $h : \{-1,1\}^n \to \{-1,1\}$. We say that A learns $\mathcal{C}$ with error ϵ if for any $f \in \mathcal{C}$, with high probability A outputs an h which satisfies $\Pr_x[f(x) \neq h(x)] \leq \epsilon$.

For boolean valued functions f and h, $\Pr_x[f(x) \neq h(x)] \leq \epsilon = \frac{1}{4}\|f-h\|_2^2$. Thus the above condition is equivalent to $\|f-h\|_2^2 \leq 4\epsilon$. The quantity $\Pr_x[f(x) \neq h(x)]$ is called the generalization error or risk in machine learning. Thus minimizing $\Pr_x[f(x) \neq h(x)]$ is equivalent to minimization of $\|f-h\|_2^2$.

We further note that Parseval's identity implies that $\|h(x)-f(x)\|^2 = \left\|\hat{h}(x)-\hat{f}(x)\right\|^2$. Thus, if we can obtain a function h having fourier coefficients close to f, then h will also be close to f in squared norm over the uniform distribution defined on the inputs. As noted above, this is sufficient to ensure that the risk $\Pr_x[f(x) \neq h(x)]$ is small.

Thus our general strategy to learn a function f would be to obtain an h whose Fourier coefficients approximate those of f.

However, the function h, obtained by approximating the Fourier coefficients may not be Boolean valued. To ensure this, we instead use the function $\text{sign}(h(x))$.

To approximate the Fourier coefficients of f, we rely on its values obtained at random samples of input points. The number of such samples, known as the query complexity will increase as the desired accuracy increases. Furthermore, the query complexity will also be large if a large number of Fourier coefficients need to be determined. Thus to allow estimation of most of the mass of Fourier coefficients, we need an efficient technique to approximate Fourier coefficients using a few samples as well as a restricted concept class of functions requiring the estimation of a only few coefficients. To summarize, it is desirable to satisfy the following two conditions:

1. Ability to estimate Fourier coefficients
2. The number of worthwile Fourier coefficients should be small.

In light of the above discussion, we restrict our concept class of functions to be ϵ-concentrated on some fixed collection of sets $\mathcal{F}$. We can then prove the following theorem:

Theorem 1. *Assume access to random examples from target function $f\colon : \{-1,1\}^n \to \{-1,1\}$. If we can identify a collection $\mathcal{F}$ of subsets such that f is $\frac{\epsilon}{2}$-concentrated on $\mathcal{F}$, then f can be learned in time poly$(|\mathcal{F}|, n, \frac{1}{\epsilon})$in addition to the time to identify $\mathcal{F}$.*

To prove the above theorem, we construct an algorithm that outputs, with high probability an h such that h is close to f.

We first make the following observation: Let f' be the restriction of f to fourier coefficients in $\mathcal{F}$ i.e $f' = \sum_{S\in\mathcal{F}} \hat{f}^2(S)$. Let g any function having all non-zero Fourier coefficients on the characters corresponding to sets in $\mathcal{F}$. We observe that:

$$\begin{aligned}\left\|\hat{g}(x)-\hat{f}(x)\right\|^2 &= \sum_{S\in[F]} (\hat{g}(x)-\hat{f}(x))^2 + \sum_{S\notin[F]} (\hat{f}(x))^2\\ &= \sum_{S\in[F]} (\hat{g}(x)-\hat{f}(x))^2 + \frac{\epsilon}{2}\\ &= \|g-f'\|^2 + \frac{\epsilon}{2}\end{aligned}$$

Thus to ensure that $\|h-f\|^2 \leq \epsilon$ it is sufficient to obtain an h satisfying $\|h-f'\|^2 \leq \frac{\epsilon}{2}$.

Our final algorithm is as follows:

```
for $S \in \mathcal{F}$ do
    Obtain an estimate of $\hat{f}(S)$ using $t$ samples of $f(x)$
    $\hat{h}(S) \leftarrow \frac{1}{t}(\sum_{i=1}^{t} f(x_t)\chi_S(x_t))$
end
$h(x) \leftarrow \sum_{S\in\mathcal{F}} \hat{h}(S)(\chi_S(x))$
Output sign($h(x)$)
```

To prove that the fourier coefficients of $h(x)$ are close to those of f with high probability, we utilize the following theorem [Vershynin, 2018], that can be derived through a Chernoff bound, i.e. by applying Markov's inequality to e^{tX} where X is a random variable.

Theorem 2. ***Hoeffding's inequality for general bounded random variables****: Let $X_1, \dots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{N}(X_i - \mathbb{E}X_i) \geq \epsilon'\right\} \leq \exp\left(-\frac{2\epsilon'^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$$

Applying the above inequality to $-X_i$ yields $\mathbb{P}\left\{\sum_{i=1}^{N}(X_i - \mathbb{E}X_i) \leq -\epsilon'\right\} \leq \exp\left(-\frac{2\epsilon'^2}{\sum_{i=1}^{N}(M_i-m_i)^2}\right)$. Thus using a union bound, we have the following inequality on the absolute deviation from mean:

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N}(X_i - \mathbb{E}X_i)\right| \geq \epsilon'\right\} \leq 2\exp\left(-\frac{2\epsilon'^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$$

We note that, for any set $S \in \mathcal{F}$, the random variable $\hat{h}(S) = \frac{1}{t}(\sum_{i=1}^{t} f(x_t)\chi_S(x_t))$ is the sum of t independent random variables $\frac{f(x_t)\chi_S(x_t)}{t}$ lying in the range $[-\frac{1}{t}, \frac{1}{t}]$ and having expectation $\hat{f}(S) = \frac{1}{t}\mathbf{E}_x[f(x)\chi_s(X)]$. Therefore, by applying Hoeffding's inequality to $\hat{h}(S)$ we obtain:

$$\mathbb{P}\left\{\left|\hat{h}(S) - \hat{f}(S)\right| \geq \epsilon'\right\} = \mathbb{P}\left\{\left|\frac{1}{t}(\sum_{i=1}^{t} f(x_t)\chi_S(x_t)) - \hat{f}(S)\right| \geq \epsilon'\right\} \leq 2\exp\left(-\frac{\epsilon'^2 t}{2}\right)$$

Thus to ensure that $\left|\hat{h}(S) - \hat{f}(S)\right| \leq \epsilon$, with probability greater than $1 - \delta'$, it is sufficient to have $2\exp\left(-\frac{\epsilon'^2 t}{2}\right) \leq \delta'$ or equivalently to sample $t \geq 2\log\frac{2}{\delta'}\frac{1}{\epsilon'^2}$ points.

Now, by choosing $\epsilon' = \sqrt{\frac{\epsilon}{2|\mathcal{F}|}}$ and $\delta' = \frac{\delta}{|\mathcal{F}|}$, and applying a union bound over all subsets $S \in \mathcal{F}$, we obtain that with probability greater than $1 - |\mathcal{F}| \times \frac{\delta}{|\mathcal{F}|} = 1 - \delta$, we have $\|h(x) - f'(x)\|^2 = \sum_{S\in\mathcal{F}}(\hat{h}(S) - \hat{f}(S))^2 \leq |\mathcal{F}| \times \frac{\epsilon}{2|\mathcal{F}|} = \frac{\epsilon}{2}$. Since the number of queries, given by $|\mathcal{F}|t$ is $poly(|\mathcal{F}|, n, \frac{1}{\epsilon})$, this proves our theorem.

Bibliography

R. O'Donnell. *Analysis of Boolean Functions.* Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139814782.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, Nov. 1984. ISSN 0001-0782. https://doi.org/10.1145/1968.1972. URL `https://doi.org/10.1145/1968.1972`.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. https://doi.org/10.1017/9781108231596.